DEPARTMENT OF COMPUTER SCIENCE AND SOFTWARE ENGINEERING

PHD COMPREHENSIVE EXAM

Service Assurance in 5G Networks

by:

Nguyen Phuc, Tran

Student ID: 40216113

PhD Committee:

Dr. Brigitte Jaumard, CSSE (Supervisor)

Dr. Rodolfo Coutinho, ECE (External Examiner)

Dr. Sandra L. Céspedes, CSSE (Examiner)

Dr. Eugene Belilovsky, CSSE (Examiner)

December 2022

Contents

1 Abstract

In this report, we are going to introduce about 5G networks and review the fundamental implementation of 5G technologies based on 3GPP specifications. Then, a brief overview of the architecture for the end-to-end network slice and its key performance indicators for monitoring the performance will be presented. Finally, the concept of service assurance will be reviewed to facilitate the effective deployment of 5G networks.

This report mostly references and summarizes these documents: System architecture for the 5G System (5GS) [1], Definitions of terms related to quality of service [2], ETSI TS 128 554 V16.7.0 (2021-01) [3] and O-RAN.WG6.CADS-v4.00 TR [4].

2 Introduction

The telecommunications industry is now one of the fastest-growing industries. And 5G, the fifth-generation technological standard, has provided several benefits that enable operators to accommodate their system expansion as well as upgrade from the current architecture. Thus, the 5G technology standard for broadband cellular networks has been studied in recent years, applied worldwide, and replaced the previous generation (4G) as of today. 5G is about significantly more than enhanced bandwidth; it also represents a fundamental redesign of the access network that leverages numerous important technological developments and places it in a direction to enable far more innovation. In other words, 5G would encourage the transition from a single access service (broadband connection) to a greater diverse array of edge services and devices.

5G is advantageous just because it not only can handle millions of high-speed devices efficiently but also because it has the potential to change global industrial innovation. On the other hand, developing and deploying 5G networks, as well as upgrading the current infrastructure, still posed some remaining challenges [5, 6, 7]. One of the challenges that operators are facing is how to ensure the quality of service which fulfill the agreements with the customers. Thus, network slicing and service assurance are potential concepts for addressing the aforementioned challenge.

In the rest of this report, we are going to review 5G architecture and its characteristics to enable end-to-end (E2E) network slicing, key performance indicators (KPIs) for network slicing and quality of service techniques in 5G networks.

3 5G architecture

3.1 Overview

Today, the cellular network offers wireless access to mobile devices (User Equipment, or UE) and these UEs may include smart devices (smartphones, tablets, household appliances), automobiles, drones, industrial (agricultural machinery, robotics) and medical equipment, among others. 5G provides wireless cell networks with high speed, enhanced reliability, and low latency via a unified, more capable air interface. It has been developed to enable next-generation customer experiences, support new deployment patterns, and offer new services. Fig. 1 illustrates a common simplified 5G network which is used by network operators. The network architecture may or may not have a ring topology, and is usually constructed with standard components.

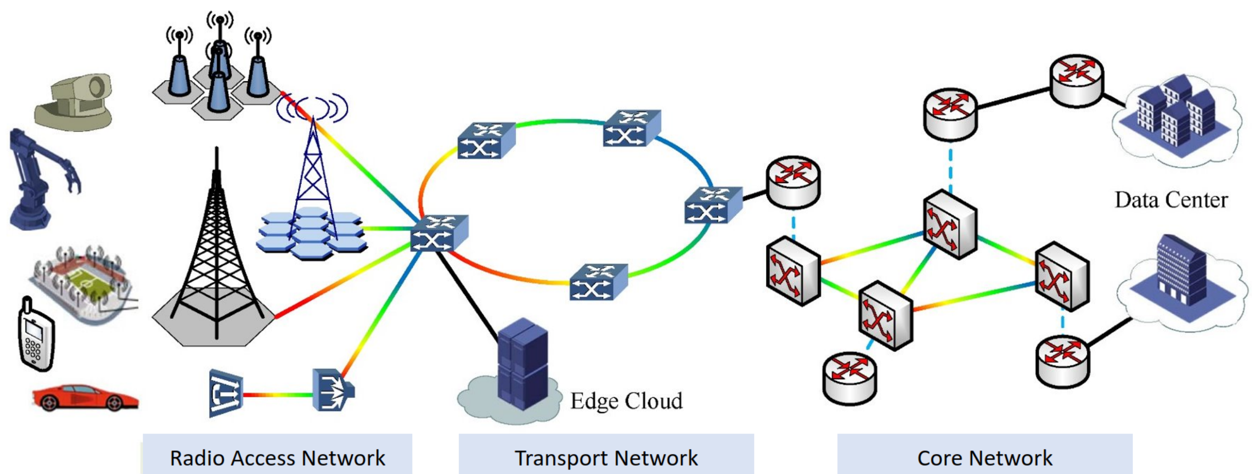

Figure 1: An overview of network architecture (extracted from [8])

Radio Access Network (RAN), Transport Network (TN), and Mobile Core Network (Core Network - CN) are often addressed when discussing the 5G network and how they operate together to enable a telecom network to fulfill the 5G technologies. Fig. 2 illustrates three common parts in a 5G network in order to enable a wireless network at a high level. The RAN refers to a group of distributed base stations or gNodeBs (gNB(s)), that maintains the radio spectrum, ensuring that it is utilized effectively and satisfies every user's quality-of-service needs. The connection services required by the 5G network are handled by the 5G transport network, which includes the access, aggregation, and core layers. A core network is the core component of a 5G network that allows access to its services and delivers multiple services to interconnected customers.

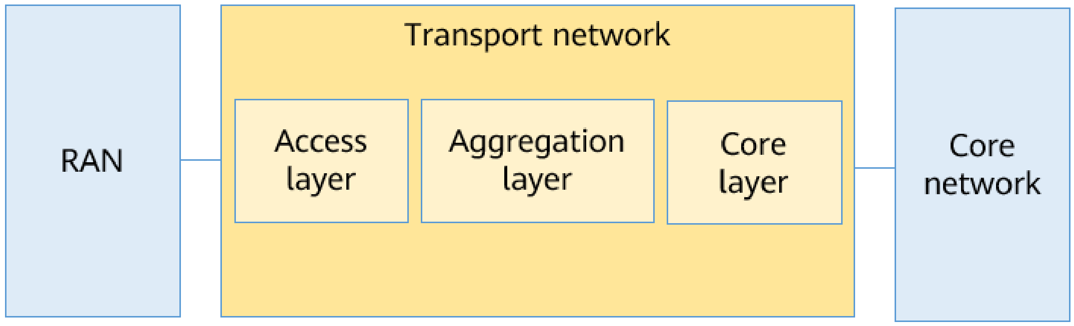

Figure 2: Common 5G cellular network consists of a RAN, TN and a CN (source [9])

The architecture of the 5G system is designed to accommodate data connection and services, which will enable the leverage of techniques such as software-defined networking (SDN), and network function virtualization (NFV). SDN is expected to promote the adaptability of networks by being characterized as a method for planning, constructing, and operating networks that separates the network's control and forwarding planes, allowing network managers to be directly programmable and the underlying infrastructure to be abstracted for applications and network services. On other hand, NFV enables the replacement of network services on specialized appliances (e.g., routers, load balancers, and firewalls) with virtualized instances operating as software on commercially available hardware. In the next sections, we are going to review three main parts of the 5G network mentioned above: RAN in section 3.2, TN in section 3.3 and CN in 3.4.

3.2 Radio access network (RAN)

In a 5G network, a RAN utilizes 5G radio FDD (Frequency Division Duplex) frequencies to offer wireless connection to UEs so they can provide incredible applications and manage wireless network resources between multiple devices in an effective way. Some 5G antennas that are integrated with radios provide a large computational capacity, including several billion transistors; enhanced performance in the lower frequency bands and in traditional radios. Throughout the 3GPP specifications, the RAN plays an essential role in providing the connection between UE and CN; it could be considered as a specialist forwarder. RAN breaks outgoing IP packets into physical layer parts and schedules their transmission over the available radio spectrum in the downstream direction. In the opposite direction, it combines physical layer segments into IP packets and transmits them upstream of the CN user plane (UP). RAN determines whether to transmit based on observations of wireless channel quality and per-subscriber rules:

- Transmit incoming packets to the UE directly.
- Transmit packets to the UE via a nearby base station indirectly.
- Apply multiple routes to reach the UE. Distributing the physical payloads across several base stations or various carrier frequencies of a single base station is an alternative in this instance.

Fig. 3 illustrated a 5G RAN architecture which could support both 5G and previous generation (4G) connections.

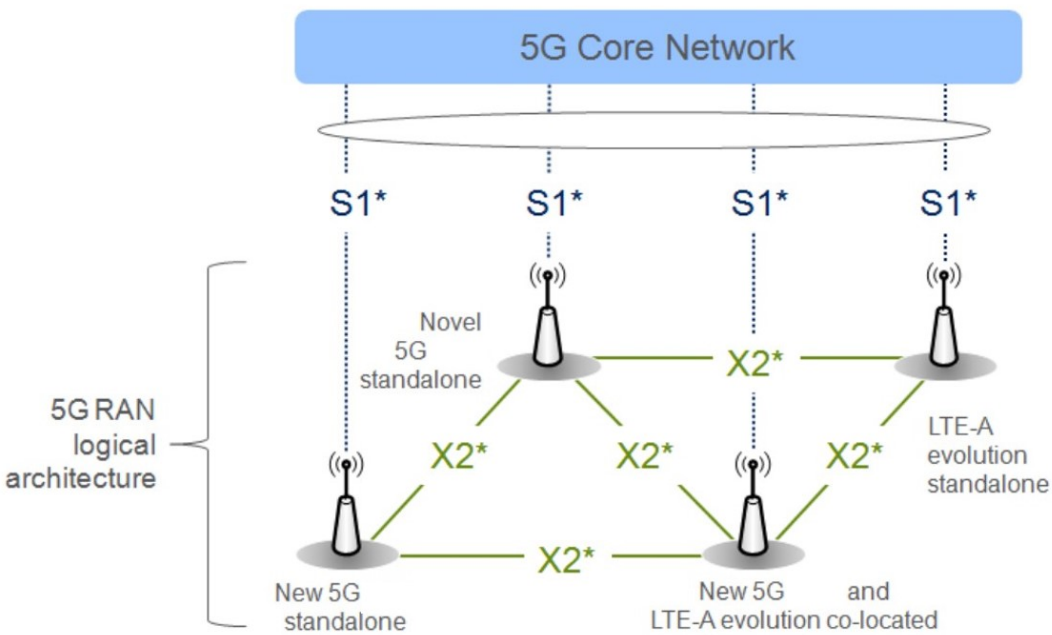

Figure 3: A reference of 5G RAN which supports both 4G and 5G (source: [10])

By implementing various virtual network functions (VNFs) for networks (e.g. virtual firewalls, virtual evolved packet cores (vEPCs), etc), 5G-RAN offers advantages to operators that have reduced expenses and established a more flexible infrastructure. With the support of VNF for the new RAN, O-RAN (or Open Radio Access Network) Alliance [4] has proposed an architect in Fig. 4, which enables shared hardware and uses more software on RAN by decoupling hardware and software. This enables a higher integration and allows incredibly capable and efficient implementations of massive MIMO (Multiple Input Multiple Output), beam shaping, and beam tracking, which are required to utilize the relatively high 5G frequencies effectively. In addition, O-RAN transfers radio processing among logical nodes, hence determining the characteristics of these connections, especially for the fronthaul. There are three different levels to consider when addressing the decoupling of hardware and software:

- The hardware layer can be found at the bottom of Fig. 4; it could be considered as a 3GPP-defined hardware sub-layer.
- An intermediary layer including Cloud Stack (e.g. containers, virtual machine (VM), etc) and acceleration abstraction layer functionality.
- A layer that provides support for virtual RAN (vRAN) functions.

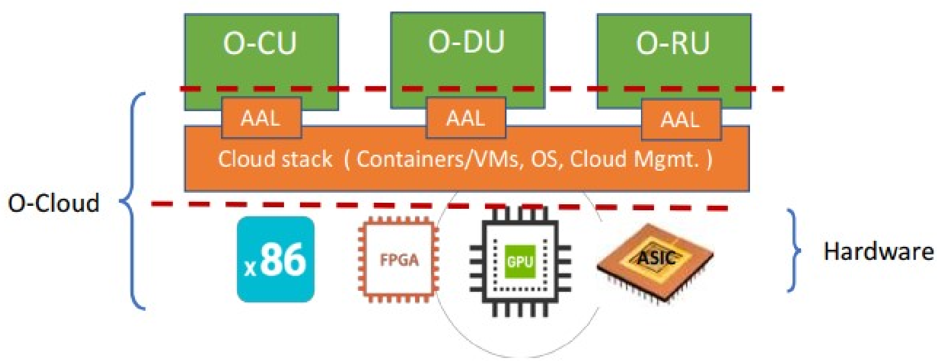

Figure 4: Decoupling RAN concept (extracted from [4])

Based on the design from Fig. 4, each layer of RAN can come from different providers including both hardware and software. In addition, it offers Cloud Platform integration for providing vRAN functionalities. The general definition of a Cloud Platform contains the following characteristics:

- The Cloud Platform is a collection of hardware and software components that enable RAN network functions to be executed via cloud computing.
- The hardware contains computation, networking, and storage components, as well as different acceleration technologies needed by RAN network functions to fulfill their performance goals. It could be shared across the system.
- The software offers well-defined, open application programming interfaces (APIs) that enable control of the whole life cycle for network operations.

- The software and hardware of the Cloud Platform also are decoupled.

3.2.1 RAN splitting

Since release 15, 3GPP established a new flexible architecture for the 5G RAN in which the base station or gNodeB (gNB) is divided into three logical nodes: the Central Unit (CU), the Distributed Unit (DU), and the Radio Unit (RU), each capable of hosting multiple roles of the 5G NR (new radio) stack. Fig. 5 depicts an example of how 5G RAN splitting in order, RU is closer to the hardware layer and often sticks with DU, then CU provides connection directly to CN.

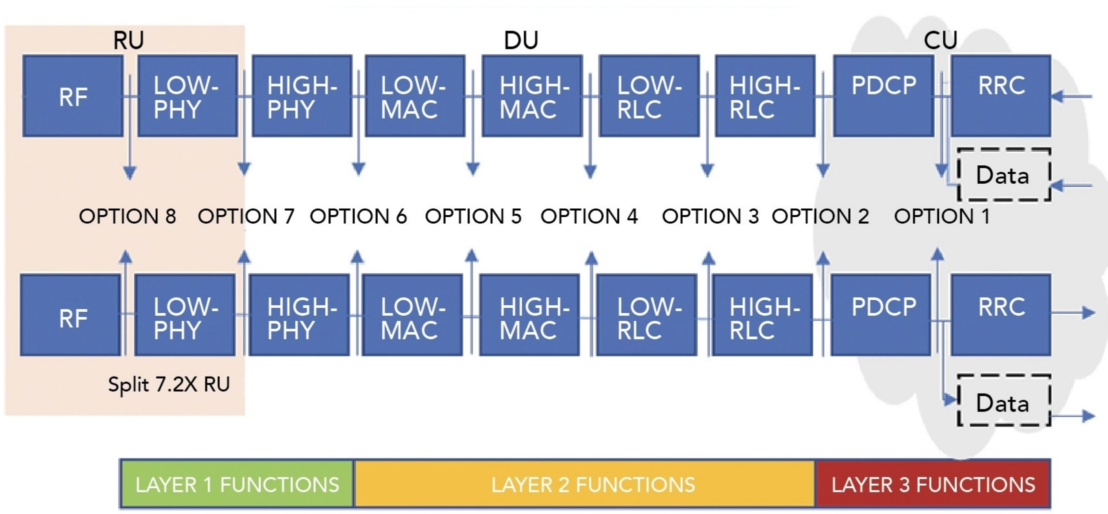

Figure 5: Functional splits define how the 5G NR Stack is allocated to the logical nodes (extracted from [11])

- RU: The radio unit is responsible for the digital front end, physical layer components, and digital beamforming functions. The primary design considerations for RUs are size, weight, and energy consumption.

- DU: the distributed unit resides next to the RU and operates the RLC, MAC, and portions of the physical layer, as well as a subset of the eNB/gNB functions (depending on the functional split option); its operation is controlled by the CU.

- CU: The centralized unit runs the RRC (Radio Resource Control) and PDCP (Packet Data Convergence Protocol) layers. The gNB is composed of a CU and one DU, with the DU being linked to the CU by a Fs-C interface for the Control Plane (CP) and a Fs-U interface for the UP. One CU can support multiple DU(s) which enables it to support multiple gNB(s).

Using the above design division, operators may do the most of network upgrades at the CU, hence the need for fewer site visits. In addition, the simplified RU enables capacity for supporting various Radio Access Technologies (RAT), thereby reducing the footprint of the remote antenna that must serve multiple cellular generations.

In the industry, when talking about fronthaul we often consider the lower level interface (eCPRI - Enhanced Common Public Radio Interface) that connects RU and DU which delivers the lowest latency at a lower cost. Using eCPRI helps fronthaul reduce latency to 100 μsec and allows a single DU to serve RUs located up to several kilometres away. The term midhaul is considered as the connection between DU and CU/MEC (multi-access edge computing). There are several options in 3GPP, but option-2 is the sole divide considered de facto between DU and CU, despite the fact that other splits are possible. The link in midhaul should have a latency of around 1 $m$sec so a centralized CU manages DU(s) within an 80-kilometre radius. Lastly, the term of backhaul refers to the connections between CU and CN and 5G CN might be up to 200 kilometres from CU.

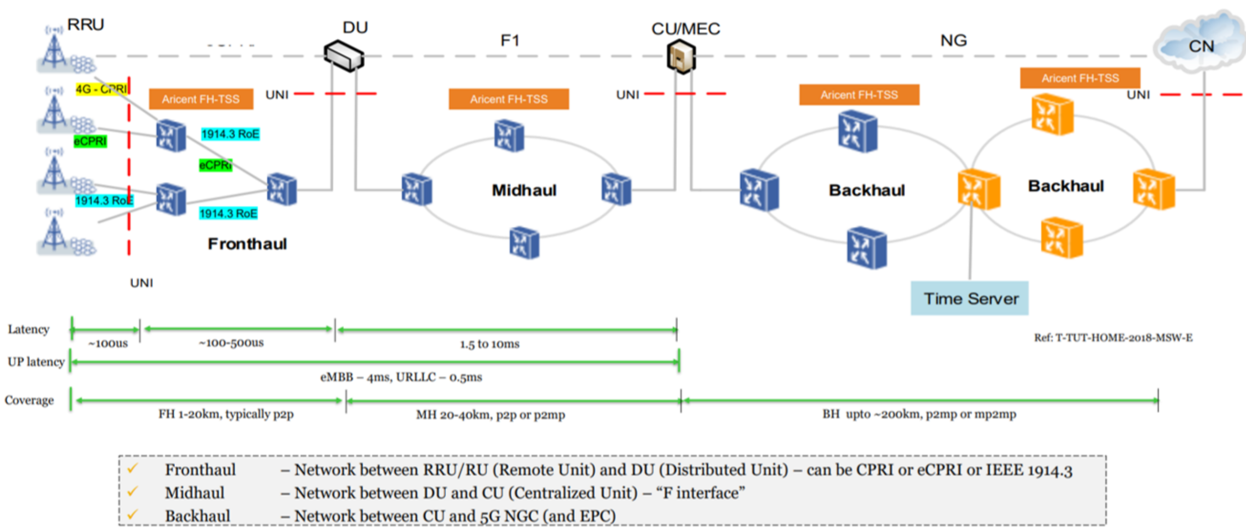

Figure 6: The interfaces between the RU, DU and CU (source [12])

Fig. 6 depicts the fronthaul, midhaul and backhaul between main components in 5G RAN as an example infrastructure. The option of how to partition NR functions inside the infrastructure is influenced by deployment scenarios, limitations, and targeted supported use cases. For example, a need to support specific QoS, support for particular user density and load demand in a specified region or existing transport networks have varying degrees of performance.

3.3 Transport network (TN)

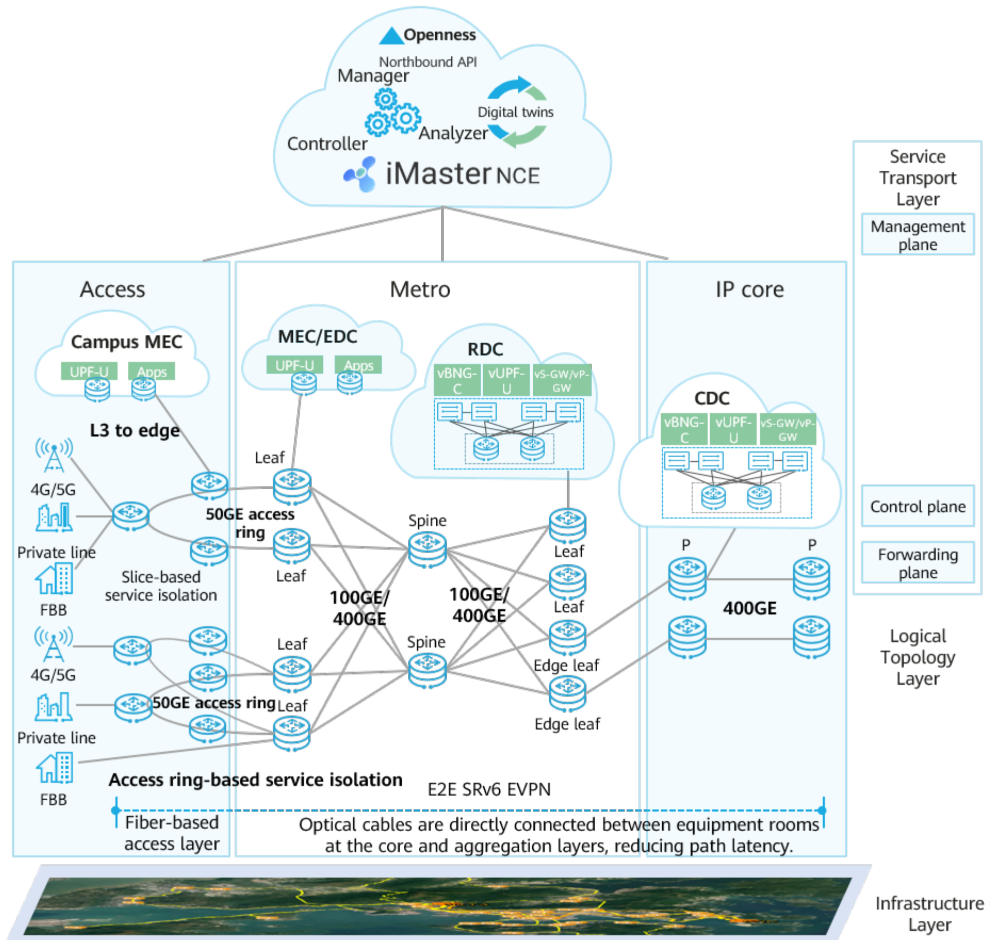

Figure 7: An example target architecture in 5G TN (source [9])

The 5G transport network (includes the aggregation, access, and core layers), illustrates in Fig. 7, provides the 5G network with connection services and the transport network must satisfy the following criteria:

- Ultra-high bandwidth: In enhanced mobile broadband (eMBB) scenarios, 5G must support a peak data rate of at least 10 to 20 Gbit/s (10 to 20 times higher than that of 4G). In massive machine-type communications (mMTTC) scenarios, the transport network must support connection densities of 1 million connections per square kilometre.

- Ultra-low latency: In eMBB and ultra-reliable low latency communications (uRLLC) scenarios, the latency on the user plane and control plane must be less than 0.5 ms (4 ms for uRLLC) and 10 ms, respectively.

- Flexible and intelligent: The transport network should support the three 5G scenario types. It must offer network slicing so that distinct situations or services have their own logical networks. Additionally, the 5G transport network must handle several comprehensive services, including 4G, 5G, and private line services.

5G imposes additional requirements on elements such as the accuracy of time synchronization, the dependability of the transport network, and its security. These needs must be taken into account throughout the development and implementation of the 5G transport network. There are three TN levels that should be considered for 5G implementation and deployment: service transport, logical topology and infrastructure.

The service transport layer contains the management plane, control protocols (in the control plane) and forwarding plane. The management plane often provides the network controllers, digital twin and simplified orchestration & management such as open interfaces, intelligence functions, service provisioning, real-time network status awareness, etc. The control protocols provide simpler network protocols for seamless connection; replace old L2VPN/L3VPN with MPLS/SRv6 and EVPN. The forwarding plane offers efficient packet forwarding in the network such as using L3 to edges to prevent detouring in the transport network.

The devices and logical connections provided by the logical topology layer offer a simplified design, ultra-high bandwidth, secure communications, and elastic scalability. In 5G and cloud-oriented transport networks, the flat spine-leaf design is used at the aggregation and core levels of the metro network and in data centres; rings are used at the access layer to provide high network resilience and scalability. In addition, the concept of network slicing may be used to separate services and offer customizable transport capacity.

The infrastructure layer offers equipment rooms and optical cables to optimal TCO (total cost of ownership) and reliable infrastructure assurance. Access layer infrastructure is designed based on a full-service fibre grid to achieve the lowest TCO, and optical fibres are used to assure highly reliable connections and deliver high bandwidth. In contrast, direct optical cable connections between equipment rooms at the core or aggregation layer guarantee low transmission distance, hence minimizing route delay.

3.4 Core network (CN)

5G core network is designed based on the service-based architect and these common services could be found in section 7, [1]. Most of the NFs in 5G is implemented by SDN and NFV technologies which are important concepts to help expand the 5G network. Fig. 8 encapsulates a common architecture and its network functions (NFs) for UEs concurrently using the reference point representation showing how various network functions interact with each other.

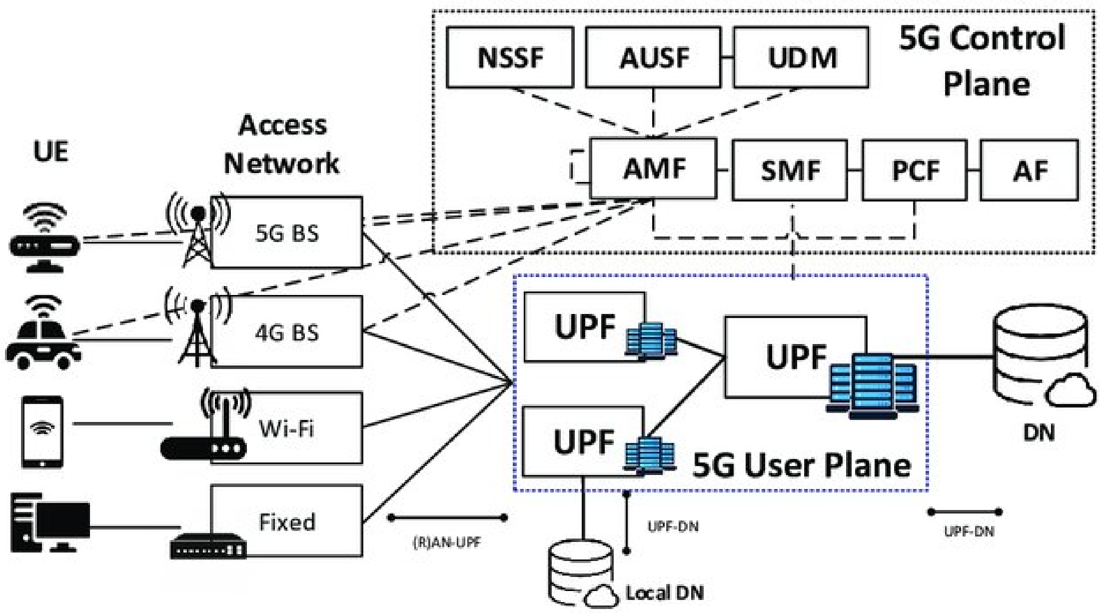

Figure 8: Common 5G architecture based on CUPS (extracted from [13])

Furthermore, CN is a collection of capabilities that fulfills several functions such as:

- Provides Internet (IP) connectivity for both data and voice services.
- Ensures that this connection meets the specified QoS specifications.
- Monitors user mobility to guarantee continuous service.
- Monitors subscriber use for invoicing and charging purposes.

The 5G CN design should include service-based interactions between CP Network Functions (NF) when indicated, also depict in Fig. 8. The following are some fundamental ideas and guidelines to follow by 3GPP:

- Decouple the User Plane (UP) from the CP to enable independent scalability, evolution, and flexible deployments, e.g. centralized or dispersed (remote) locations. UP and CP are similar to some documents that are control plane and data plane or in 3GPP is CUPS, Control and User Plane Separation.
- The design of the NF should be modularized to allow for network slicing that is flexible and efficient.
- Define procedures, also known as the collection of interactions that occur between network functions, as services whenever it is possible to do so, in order to make their reuse possible.
- Enable each NF and its Network Function Services (NFS) to communicate with other NFs and their NFS either directly or indirectly by using a Service Communication Proxy. The design does not prevent the utilization of an additional intermediate function in order to assist in the routing of CP messages (e.g. DRA).
- Minimize dependencies that exist between the RAN and the CN. The design should support different access types, such as 3GPP access and non-3GPP access, with a converged core network by a common interface between AN and CN.
- Provide a uniform authentication infrastructure.
- Support "stateless" NFs in which the computation resource is divorced from the storage resource.
- Support capability exposure.
- Allow simultaneous access to both local and centralized services. UP functions may be placed close to the AN with the purpose of providing low-latency services and local data network access.
- Support roaming for both Home-routed and Local breakout traffic inside the visiting public land mobile network (PLMN).

Based on the reference design, there are two methods to describe the interaction between NFs:

- A service-based representation in which network functions (e.g. AMF) in the Control Plane allow other authorized NFs to access their services. This illustration also provides point-to-point (P2P) references when applicable.
- A reference point representation demonstrates the interaction between NFS and NF defined by a P2P reference point (e.g. N11) between any two network functions (e.g. AMF and SMF).

CN is an essential part of 5G, offering a bridge between the RAN in a particular geographic area and the global IP-based Internet. The 3GPP 5G specification has enabled great deployment flexibility for the CN, e.g. each instance of the CN serves a city area. In the next part, we will explain CP and UP, the two most significant components of CN, in order to explain why CN's deployment is flexible.

3.5 User plane and control plane

In reference to 3GPP 5G System (5GS), CUPS decouples control and user plane functions to enable decentralization of the data forwarding component. This supports packet processing and traffic aggregation to be handled closer to the network edge, hence enhancing network performance. The major objective of CUPS was to promote 5G New Radio installations to enable early Internet of Things (IoT) applications and faster data speeds.

For more details, CP refers to all signalling utilized to support the 5G network system operations that establish and maintain the UP. In addition, the CP is a forwarding channel for the exchange of service-related information that needs to be reliable, scalable and efficient to meet the requirements of mobile network operators. And, CP contacts connected with each registered UE occur as required inside the Core Network. Therefore, it is essential that the control plane interactions occur effectively.

CP plays the most crucial role in 5G CN operations, including mobility management, access control, data packet routing and forwarding, radio resource management, and UE reachability. When the UE connects to the RAN, whether during power-up or handover, the base station will establish a 3GPP CP connection between the UE and the relevant CN CP component as well as forward signalling traffic between them. Then, the base station constructs one or more tunnels for each active UE between the associated CN UP components. Next, the base station passes both CP and UP packets, which are tunnelled across SCTP/IP and GTP/UDP/IP, between the CN and UE. In addition, the base station handles UE handover with neighbouring base stations through direct station-to-station interconnections that are identical to station-to-CN communication.

Following the organization of the set of functional blocks described in the previous section, these components below could be considered to belong to the CP group:

- Core Access and Mobility Management Function (AMF): In charge of managing connectivity and reachability, mobility, access authentication and authorization, and navigation services.
- Session Management Function (SMF): Handles every UE session, including IP address assignment, selecting of related UP function, QoS management, and UP routing authority.
- Policy Control Function (PCF): Handles the policy rules that are then enforced by other CP services
- Unified Data Management (UDM): Controls user identification via providing access permission.
- Authentication Server Function (AUSF): it is a server for authentication basically.
- Structured Data Storage Network Function (SDSF): A service used to store structured data and could be achieved using a ”SQL Database” in a system based on microservices.
- Unstructured Data Storage Network Function (UDSF): A service used to store unstructured data. Could be achieved using a ”Key/Value Store” in a system based on microservices.
- Network Exposure Function (NEF): A method for exposing chosen capabilities to third-party services, including data representation transition between internal and external formats. It could be accomplished by an ”API Server” in a system based on microservices.
- NF Repository Function (NRF): A method for identifying available services that could be achieved by a ”Discovery Service” in a system based on microservices.
- Network Slicing Selector Function (NSSF): A method for determining a network slice to serve a certain UE.

The second group could be known as UP and mostly is User Plane Function (UPF) in 5G systems. The goal of the UPF is to forward traffic between the RAN and the Internet. It is also responsible for policy enforcement, legal intercept, traffic consumption reporting, and QoS policing, in addition to packet forwarding. In detail, UPF provides the packet processing framework for Service-Based Architectures (SBAs). According to 3GPP, UPF provides:

- The interface between the mobile infrastructure and the Data Network (DN), consisting of the encapsulation and decapsulation of GPRS Tunnelling Protocol for the user plane (GTPU).
- The Protocol Data Unit (PDU) session station for enabling mobility between Radio Access Technologies (RATs), includes the transmission of one or more end indicator packets to the gNB.
- Packet forwarding and routing, including serving as an Uplink Classifier (directing flows to particular data services using traffic matching criteria) and a Branching point when serving as an Intermediate UPF (I-UPF) multi-homed to several PDU session points (PSA).
- Application identification utilizing Service Data Flow (SDF) traffic filter patterns or Packet Flow Description (PFD) provided by the SMF.
- QoS flow handling includes rate limitation and reflective QoS (DSCP) labelling on the downlink.
- Traffic consumption monitoring for charging and the interface for the Lawful Intercept (LI) collector.

UPF is a significant new element for enabling the next generation of service-based architectures. However, the improvement in this domain should also enable the measurement of the likelihood of potential CUPS customers choosing the UPF over temporary PGW/SGW-U solutions, hence easing the transition from 4G to 5G.

3.6 Security and mobility

When discussing the security of 5G networks, two trust principles are considered:

- Base station trusts that it is connected to the CN: It's trusted by a secure private network in an established tunnel when via a GTP/UDP/IP tunnel to the Core-UP and an SCTP/IP tunnel to the Core-CP.

- The trust of the subscriber between UE and the operator: Each UE has a unique identifier (often provided by a SIM card, or phone number and includes a secret key used for authenticating itself) and establishes the radio parameters (i.e. frequency band) needed to communicate with that operator's base stations.

Fig. 9 illustrates the sequence for establishing a secure connection between UE and CN. Firstly, a connection between UE and the nearest base station over a temporary radio link when a UE becomes active (1). The base station then forwards the packet for authentication to the Core-CP and Core-UP throughout the existing tunnel (between the base station and CN) (2). CN will determine and identify a collection of authentication and encryption alternatives, with the actual protocols (e.g., RSA, AES) being an implementation decision. When the Core-CP and UE have confirmed each other's identities, the Core-CP notifies the other components of the UE service contexts, e.g. setup CP (assign an IP, set QoS parameter), (3). Finally, an end-to-end UP channel will be formed through Core-UP, and UE can also be utilized (4).

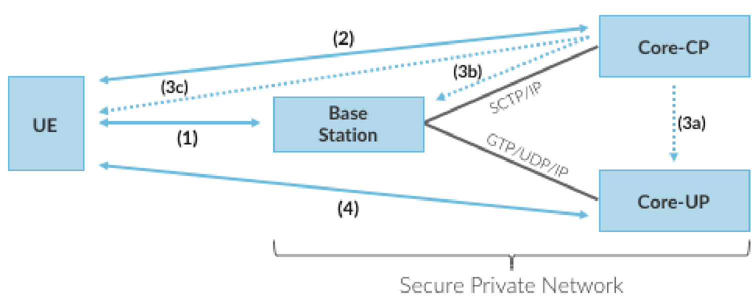

Figure 9: Sequence of steps to establish secure CP and UP channels (extracted from [14])

In addition to security procedures which may affect the Service assurance (SA), the UE's movement throughout the network may also be a factor. Mobility support is now defined as the process of repeating one or more steps from the security procedure when a UE moves around RAN(s). When a UE moves closer to another base station, the base stations exchange the UE context based on the CQI (Channel Quality Information) and initiate the handover procedure. After the changeover is signalled, CN will reactivate the setup process from step (3) for reconstructing the UP tunnel between the UE and relevant network components as well as maintaining the UE session. When a handover operation is performed, the UPF will buffer packets during this period to prevent packet loss; nevertheless, this may increase transmission delay in the E2E network slice.

4 End-to-end (E2E) network slicing

4.1 Slicing concept

It is anticipated that 5G will offer comprehensive user interfaces, such as virtual and augmented reality (AR/VR), operation applications such as public safety, autonomous cars, and IoT applications. These scenarios will cover everything from home appliances to robotic systems to self-driving automobiles, and 5G not just enable individuals from accessing the Internet through their smart devices (smartphones, smart watches, tablets, etc) but also a lot of autonomous equipment operating on their behalf.

Supporting these services involves more than just enhancing the bandwidth or latency of customers. Thus, the slicing concept in 5G could help operators fulfill the design which meets the customer's needs. It is a logically divided, standalone, secured and self-contained network that serves distinct services with differing needs for speed, latency, and reliability. By using SDN, NFV, orchestration, analytics, and automation, operators may rapidly construct network slices to serve a particular application, service, user group, or network. SDN is a method that uses software-based controllers or APIs to interact with a network's underlying hardware architecture and manage network traffic. It is extremely important in 5G because it supports the prospective concepts: higher management with improved speed and flexibility, configurable network architecture, and strong security. NFV is essential to the support of SDN because it allows virtualized network services (e.g. routers, firewalls, load balancers, etc) that are packaged as virtual machines on commodity hardware. With the assistance of NFV, operators may increase scalability and agility by enabling service providers to deploy new network services and applications on demand without needing extra hardware resources; operators also no longer need to have specialized hardware for each network function.

On the basis of these factors, a fundamental network design is required which is demonstrated by three groups of characteristics and their use cases (see Fig. 10):

- Enhanced Mobile Broadband (eMBB): extreme data rates (e.g. multi-Gbps peak, 100+ Mbps sustained) and heavy capacity (e.g. 10 Tbps / km^2). We might consider HD video streaming as an example.

- Massive Machine Type Communications (mMTC): offers connections to a huge number of devices that transfer limited amounts of information frequently. Smart cities and industry 4.0 are representatives of this group.

- Ultra-Reliable Low Latency Communication (uRLLC): provide a guaranteed connection (over 99.999%), extreme mobility (up to 100 km/h), and ultra-low latency (approximately 1 ms) for mission-critical applications. Autonomous driving, for instance, would need such a link due to the considerable danger involved.

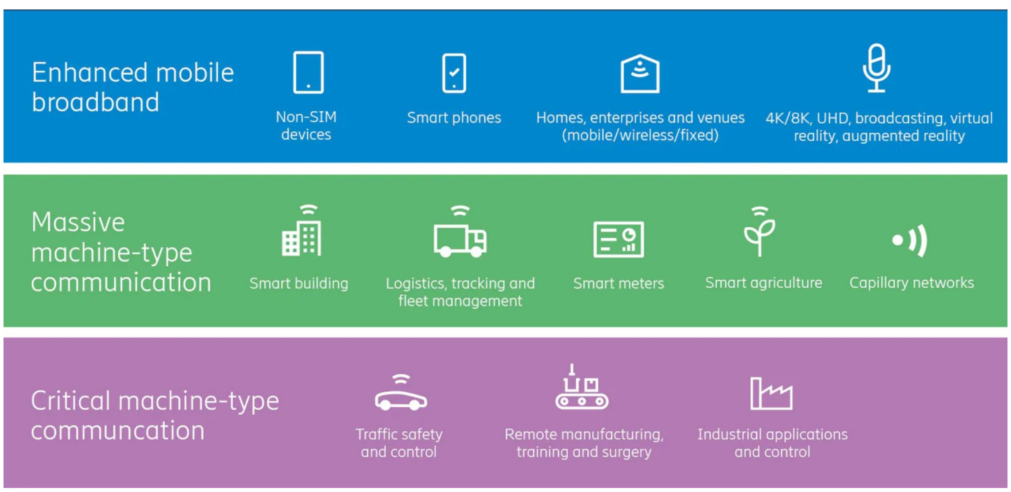

Figure 10: Common 5G use cases (extracted from [15])

In the following parts, we will review the E2E network architecture that helps meet the 5G requirements mentioned above.

4.2 E2E network slicing design

Network slicing is a significant feature of 5G that enables end-to-end connection and data processing that are adapted to particular business demands. Thus, a modern Operation Support System (OSS) and Business Support System (BSS) with automated strategy and operational procedures are required to effectively manage network slices and optimize costs. Matching to 3GPP, a network slice is a realization of the QoS Class Identifier and a network E2E slice instance is constructed inside a PLMN and combines CN CP and UP capabilities. An E2E slice may include specific NG-RAN(s) or could be shared across the Fronthaul.

According to 3GPP, there are 4 Standardized Slice Type (SST) values which could be found in [1] section 5.15.2.2. Each E2E slice may support different features (for example uRLLC, eMBB) and be defined based on applications and customers' needs. Through programmable in the scalable 5G networks, Service Level Agreement (SLA) driven orchestration and powerful AI (Artificial Intelligence), the specific network functionalities can be constructed, deployed, and controlled automatically throughout the network's life cycle. In the deployment phase, the operator may deploy one or multiple instances of a network slice with the same features and serve different UE groups. Despite the access type, the network may dynamically provide a single UE with one or more network slice instances through 5G-AN (s).

When a UE initially registers at the AMF component, the selection of the collection of network slice instances will be performed by interacting with the Network Slice Selection Function (NSSF). Then, a PDU Session inside a PLMN will be formed for each network slice instance and will not be shared with other network slice instances, even if various network slice instances may have slice specific PDU Sessions utilizing the same Data Network Name (DNN).

By decoupling hardware and software, the slicing approach could be successfully implemented at the RAN level with the benefit of virtualization. Fig 4 illustrates the decoupled RAN approach which might assist to enhance RAN flexibility and deployment velocity, as well as reducing the time and operating costs. Thus, RANs might be utilized to share across network slice instances (both center and edge networks) while still ensuring network slice isolation and respect to the KPI requirements. The purpose of RAN in network slicing is to enable programmable virtual RAN nodes (base stations) that can perform on the same hardware and share spectrum resources according to predefined rules for various services, applications, UEs, etc.

Besides slicing the RAN, we must also apply the slicing concept to CN and TN. It may have different ways to implement but the well-understood scenario requires QoS mechanisms in the network switches (enable transfer packets through the switching fabric relying on the bandwidth assigned to each slice) and the cluster processors (ensure containers/VMs implementing each VNF are assigned enough resources, such as CPU and RAM, to preserve the packet forwarding rate of the associated slice). Thus, one network slice may have a set of VNFs (e.g. UPF, SMF, AMF, etc) established as a service graph and could support scaling independently.

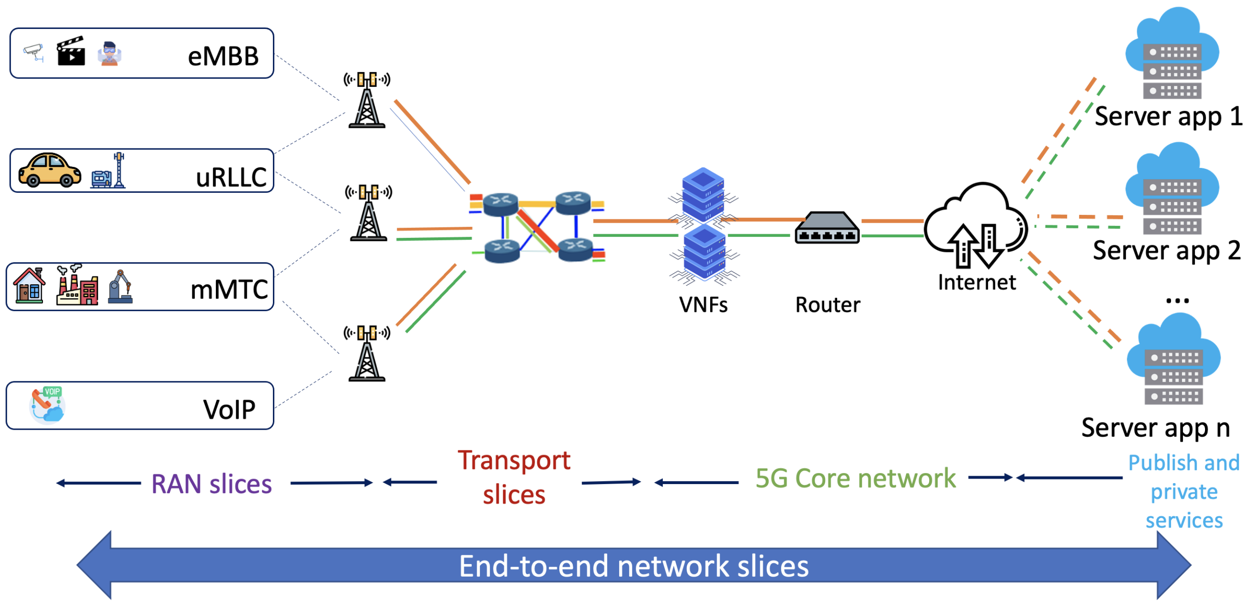

Figure 11: An example E2E slicing network topology

We could construct an E2E slicing network by linking the two previous concepts in the same network as illustrated in Fig. 11 which supports both RAN, TN and CN. Network slicing is not a new concept today as it has been successfully implemented in 5G and continues to be developed [16], Fig. 12 is an illustration of the network slicing industry velocity, and details each network slicing version. Since late 2020, Network Slicing 3.0 has been developed with the assistance of AI and is still being enhanced phase by various AI technologies [17, 18, 19]. Network slicing 3.0 requires adaptive optimization and dynamic closed-loop management of SLA for service assurance to be successful.

Consequently, to better serve businesses and industries that place a high requirement on mobility, roaming, and service continuity, Machine Learning (ML)/AI combined with a system for negative feedback can fulfill these criteria [20]. Thus, closed-loop control is also a current hot topic in 5G networks today.

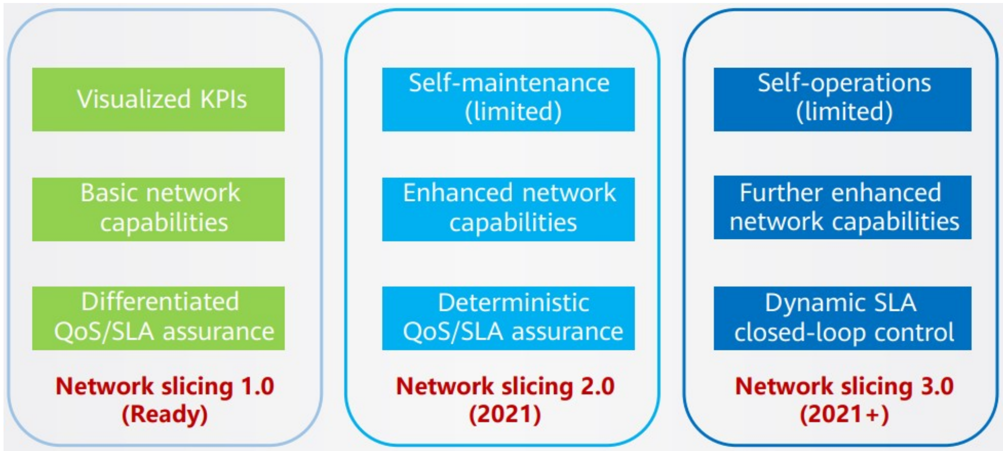

Figure 12: Network slicing industry pace (extracted from [16])

5 Service assurance (SA)

5.1 Service-Level Agreement (SLA)

In networking, an SLA outlines the products or services to be provided, the single point of contact for end-user issues, and the metrics by which the process's efficacy is reviewed and approved. SLAs assist providers in managing customer expectations and defining the severity levels and conditions under which they are not responsible for outages or performance problems. SLAs often comprise a variety of components, ranging from the definition of services through the termination of an agreement, and are also specified at several levels:

- Customer-based SLA: A contract with a specific client group that covers all the services they utilize.
- Service-based SLA: A contract applicable to all clients utilizing the service provider's services.
- Multilevel SLA: A contract consists of corporate-level, customer-level, and service-level SLAs which are used to address distinct groups of consumers for the same services.

An SLA is often defined based on the need of customers and available infrastructure from the network providers (e.g. requires a small latency, high bandwidth in several applications, etc). After a consumer has signed a contract with a network service provider, the SLAs will be converted to the QoS parameters in the network. Fig. 13 depicts an example of the SLA in a 5G network in some scenarios.

	Scenario	Experienced data rate (DL)	Experienced data rate (UL)	Area traffic capacity (DL)	Area traffic capacity (UL)	Overall user density	Activity factor	UE speed	Coverage
1	Urban macro	50 Mbps	25 Mbps	100 Gbps/km^2 (note 4)	50 Gbps/km^2 (note 4)	10 000/km^2	20%	Pedestrians and users in vehicles (up to 120 km/h	Full network (note 1)
2	Rural macro	50 Mbps	25 Mbps	1 Gbps/km^2 (note 4)	500 Mbps/km^2 (note 4)	100/km^2	20%	Pedestrians and users in vehicles (up to 120 km/h	Full network (note 1)
3	Indoor hotspot	1 Gbps	500 Mbps	15 Tbps/km^2	2 Tbps/km^2	250 000/km^2	note 2	Pedestrians	Office and residential (note 2) (note 3)
4	Broadband access in a crowd	25 Mbps	50 Mbps	[3,75] Tbps/km^2	[7,5] Tbps/km^2	[500 000]/km^2	30%	Pedestrians	Confined area
5	Dense urban	300 Mbps	50 Mbps	750 Gbps/km^2 (note 4)	125 Gbps/km^2 (note 4)	25 000/km^2	10%	Pedestrians and users in vehicles (up to 60 km/h)	Downtown (note 1)
6	Broadcast-like services	Maximum 200 Mbps (per TV channel)	N/A or modest (e.g. 500 kbps per user)	N/A	N/A	[15] TV channels of [20 Mbps] on one carrier	N/A	Stationary users, pedestrians and users in vehicles (up to 500 km/h)	Full network (note 1)
7	High-speed train	50 Mbps	25 Mbps	15 Gbps/train	7,5 Gbps/train	1 000/train	30%	Users in trains (up to 500 km/h)	Along railways (note 1)
8	High-speed vehicle	50 Mbps	25 Mbps	[100] Gbps/km^2	[50] Gbps/km^2	4 000/km^2	50%	Users in vehicles (up to 250 km/h)	Along roads (note 1)
9	Airplanes connectivity	15 Mbps	7,5 Mbps	1,2 Gbps/plane	600 Mbps/plane	400/plane	20%	Users in airplanes (up to 1 000 km/h)	(note 1)

NOTE 1: For users in vehicles, the UE can be connected to the network directly, or via an on-board moving base station.
NOTE 2: A certain traffic mix is assumed; only some users use services that require the highest data rates [2].
NOTE 3: For interactive audio and video services, for example, virtual meetings, the required two-way end-to-end latency (UL and DL) is 2-4 ms while the corresponding experienced data rate needs to be up to 8K 3D video [300 Mbps] in uplink and downlink.
NOTE 4: These values are derived based on overall user density. Detailed information can be found in [10].
NOTE 5: All the values in this table are targeted values and not strict requirements.

Figure 13: An example of performance requirements for high data rate and traffic density scenarios (source [21])

Different situations need 5G technology to accommodate very high data speeds or traffic density. The scenarios include many service locations, including urban and rural locales, the workplace and the home, and special deployments (e.g. massive gatherings, broadcast, residential, and high-speed vehicles)

5.2 Quality of Service (QoS)

Service assurance is essential in 5G networks since it helps reduce unnecessary expenditure capacity, enables new revenue models, improves cost effectiveness, and enhances the user experience. SA uses rules and procedures by a Communications Service Provider (CSP) to guarantee that services offered across networks match a predefined service quality standard to maximize the quality of user experiences (QoE) via the Quality of service (QoS) model. The 5G QoS model relies on the QoS Flow pattern and allows QoS flows that need a guaranteed flow bit rate, as well as QoS flows that do not; it also provides reflective QoS. A QoS Flow is managed by a Session Management Function (SMF) and may be prepared or set up during the PDU Session Establishment process. It contains one or more UL and DL Packet Detection Rules (PDR(s)) provided by the SMF to the UPF. Thus, the QoS Flow is the smallest granularity of QoS differentiation in the Protocol Data Unit (PDU) Session and it includes a QoS Flow ID (QFI) used to identify a QoS Flow in the 5G System. Consequently, QFI should be unique inside a PDU Session in order for UP traffic with the same QFI in a PDU Session to get the same traffic forwarding treatment (e.g. scheduling, admission threshold).

In SA, the QoS profile is used to determine whether a QoS Flow has Guaranteed Bit Rate (GBR) or has Non-GBR. The QoS parameters that are described below are included in the QoS profile that is transmitted to the RAN as part of a QoS Flow:

(1) For each QoS Flow, the QoS profile could include a 5G QoS Identifier (5QI) and allocation and Retention Priority (ARP) as parameters. The standardized 5QI to QoS characteristics mapping could be found in part 5.7.4 [1].

(2) For each Non-GBR QoS Flow only, the QoS profile may also include Reflective QoS Attribute (RQA) parameter only.

(3) For each GBR QoS Flow only, the QoS profile shall also include Guaranteed Flow Bit Rate (GFBR) - UL and DL; and Maximum Flow Bit Rate (MFBR) - UL and DL. It may also include notification control and maximum packet loss rate - UL and DL.

Figure 14 depicts the classification and identification of User Plane traffic, as well as the mapping of QoS Flows to AN resources. Uplink (UL) and Downlink (DL) are the two primary flows we should consider in the system.

In DL, the UPF classifies incoming data packets according to the Packet Filter Sets of the DL PDRs in the order of their priority. The UPF provides the category of User Plane traffic corresponding to a QoS Flow through an N3 (and N9) User Plane marking through a QFI. In

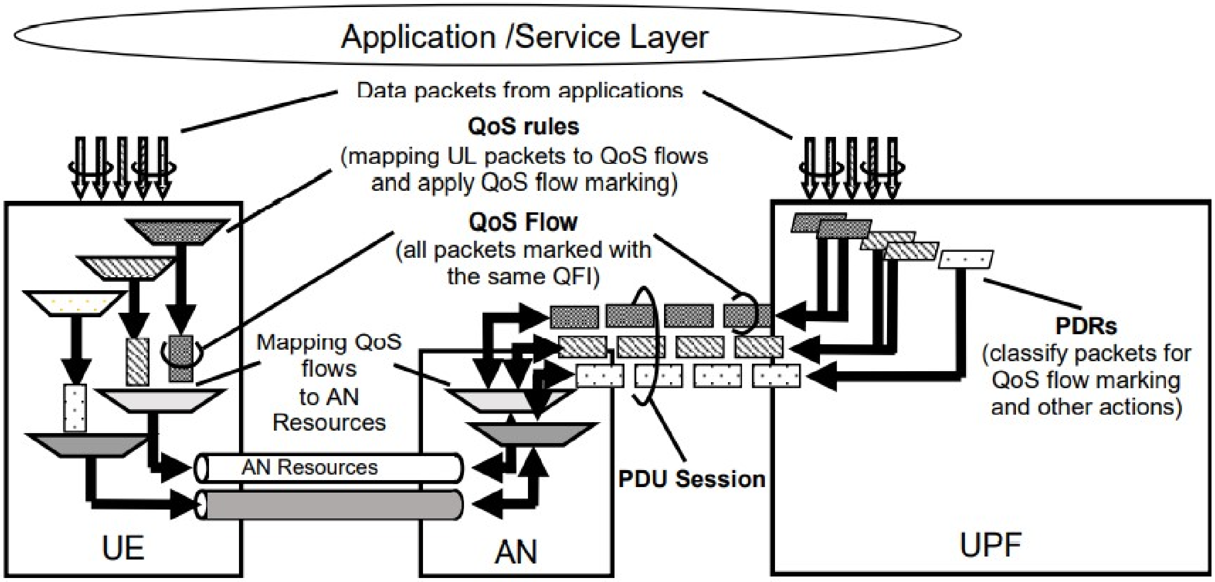

Figure 14: Classification and identification of User Plane traffic (extracted from [1])

another hand, there is no one-to-one correspondence between QoS Flows and AN resources. It is the responsibility of the AN to identify, make available and release the required AN resources to which QoS Flows may be mapped. When the AN resources onto which a QoS Flow is mapped are released, the AN will notify the SMF and if there is no matching DL PDR, the UPF should reject the DL data packet.

In the UL flow, the default QoS rule for an unstructured PDU Session does not include a packet filter set and accepts all UL packets. In contrast, during a PDU Session of type IP or Ethernet, the UE verifies UL packets using UL Packet Filters in the Packet Filter. Then, set in the QoS rules depending on the priority value of QoS rules in ascending order until a matching QoS rule (whose packet filter matches the UL packet) is discovered. Similarly to DL, UE should reject the UL data packet if no matching QoS rule is identified for UL.

When connecting to the network, the UE derives QoS rules upon receiving the DL packet for UL Packet Filter within a PDU Session by the protocol ID/ next header. The protocol ID/ next header could be UDP, TCP or ESP (Encapsulating Security Payload) and is determined by using the source and destination IP addresses, the Security Parameter Index, source and destination port numbers, and the Protocol ID / Next Header field itself.

In the 5G network, SA must evaluate both service performance and customer experiences across the network components for helping operators to grow their business footprint in order to fulfill the required network KPIs. To be effective, SA may interact with various network components (e.g., UPF, orchestration management) to offer an automation mechanism (e.g., a closed loop algorithm) for auto-scaling in the network based on network KPIs and energy efficiency.

5.3 E2E network KPIs

In 5G E2E network slicing, depending on the purpose of measurement, we will have different kinds of KPIs. There are six main categories of KPIs that are used to measure network performance: Accessibility KPI, Integrity KPI, Utilization KPI, Retainability KPI, Mobility KPI and Energy Efficiency (EE) KPI.

5.3.1 Accessibility KPI

This metric is intended to determine whether services requested by users can be accessible under a certain set of circumstances; it also relates to the quality of being available when consumers need it, e.g. a user requests access to a network, phone calls, and data calls. These KPIs could be described and calculated by the following (all the formulas could be found in [3] section 6.2):

(1) Registered Subscribers of Single Network Slice Instance through AMF: describes the average number of subscribers enrolled to a network slice instance and could be calculated by summing AMF subscribers in a network slice instance.

(2) Registered Subscribers of Network and Network Slice Instance through UDM: indicates the total number of network subscribers enrolled using UDM and corresponds to the metric RM.RegisteredSubUDMNbrMean, which measures the number of subscribers registered in UDM.

(3) Registration success rate of one single network slice instance: reflects the ratio of the number of successfully completed registration processes to the number of attempted registration procedures for the AMF set belonging to a single network slice and use to evaluate the accessibility and network performance. It is calculated by dividing successful registration operations by those that were attempted.

(4) Data Radio Bearer (DRB) Accessibility for UE services: provides the DRB's setup success rate, including the RRC connection (DRB-SuC) and NG signalling connection (NG-SuC) setup success rates. The KPI value is calculated by multiplying DRB-SC by NG-SuC, which are mentioned previously.

(5) PDU session establishment success rate of one network slice (SNSSAI): describes the ratio of the number of successful PDU session establishment requests (PDU-SecR) to the number of PDU session establishment request attempts (PDU-SecRA) for the SMF within a network slice (SNSSAI) and used to evaluate the accessibility and network performance. It is calculated simply by dividing PDU-SecR by PDU-SecRA.

5.3.2 Integrity KPI

Integrity is the attribute of data that has not been modified in an unauthorized manner. Service integrity is the amount to which a service is delivered without severe impairments after it has been received. This metric is used to measure the integrity or character of the network to its consumers, such as the latency and throughput. The following KPIs belong to this category (all the formulas could be found in [3] section 6.3):

(1) Downlink latency in gNB-DU: explains the gNB-DU component of the user's experience on packet transmission latency. It is the average time between the receiving of an IP packet by the gNB-DU and the transmission of the first part of that packet over the air interface, for a packet arriving while there is no prior data in a queue for transmission to the UE.

(2) Uplink throughput for network and network slice instance: explains the upstream throughput of a network slice instance by calculating the packet size for every successfully transmitted UL IP packet across the slice instance during each period and is used to monitor the efficiency of the integrity of the E2E network slice instance.

(3) Downlink throughput for network and network slice instance: Measures the downstream throughput of a network slice instance by computing the payload size for each successfully delivered DL IP packet across the slice instance during each time interval and is used to assess the E2E network slice instance's integrity performance.

(4) Uplink throughput at N3 interface: shows the amount of octets of all arriving GTP data packets on the N3 interface (recorded at UPF) generated by the GTP-U protocol unit on the N3 interface within a period and is used to verify upstream GTP throughput integrity at the N3 interface.

(5) Downlink throughput at N3 interface: identifies the total amount of octets of all downstream GTP data packets generated by the GTP-U protocol unit on the N3 interface (delivered downstream from UPF) over a period and used to test the performance of interface N3 integrity.

(6) RAN UE throughput: indicates how NG-RAN affects the service quality offered to an end-user. UE perceived throughput in NG-RAN is a critical performance metric for 5G network operation. For example, when the UE throughput of the NR cell doesn't fulfill the quality requirements, network actions such as reconfiguration or capacity expansion must be performed. The UE throughput metric also includes "NR option 3" cases.

5.3.3 Utilization KPI

It is essential to assess the current consumption of virtualized resources (such as memory and storage utilization) consumed by a network slice instance. If the usage exceeds or falls below the threshold, the management system may perform scaling in/out actions. The following KPIs belong to this category (all the formulas could be found in [3] section 6.4):

(1) Mean number of PDU sessions of network and network slice instances: represents the average number of PDU sessions established successfully in a network slice and is collected via the PDU session setup operations of SMFs, which are associated with the network slice.

(2) Virtualized resource utilization of network slice instance: measures the use of virtualized resources (e.g., CPU, memory, storage) assigned to a network slice and is calculated by dividing the use of virtualized resources by the system capacity delivered to the network slice.

(3) PDU session establishment time of network slice: indicates the time necessary to successfully establish a PDU session on a network slice and is used to analyze network slice usage and network performance. It is determined by monitoring the interval between SMF's receipt from AMF "Nsmf_PDUSession_UpdateSMContext Request", which includes N2 SM information

received from RAN to the SMF and the sending of a "Nsmf_PDUSession_CreateSMContext Request or Nsmf_PDUSession_UpdateSMContext Request PDU Session Establishment Request" message from AMF to the SMF.

(4) Mean number of successful periodic registration updates of a network slice: represents the average number of successful cyclical registration updates at an AMF in a network slice and is determined by accumulating the number of successful periodic registration updates at the AMFs that are associated with the network slice following registration acceptance by the AMF to the UE that issued the periodic registration update request.

5.3.4 Retainability KPI

This KPI is used to evaluate the network's ability to maintain users' possessions or offer services to users. The following KPIs belong to this category (all the formulas could be found in [3] section 6.5):

(1) QoS flow retainability: indicates the frequency with which an end-user unexpectedly loses a QoS flow during the period the QoS flow is being utilized and obtained by the number of QoS flows with data in a buffer that was abnormally released, normalized with the number of data session time units.

(2) DRB Retainability: indicates the frequency with which an end-user unexpectedly loses a DRB while the DRB is running and is determined by the amount of DRBs that were released unexpectedly and were active at the moment of their release, normalized by the number of data session time units.

5.3.5 Mobility KPI

Mobility is one of the essential characteristics that enables continuous service for mobile subscribers inside a certain area. This metric is mostly related to the handover process and its performance in the 5G network. The following KPIs belong to this category (all the formulas could be found in [3] section 6.6):

(1) NG-RAN handover success rate: indicates the frequency with which a handover inside NR-RAN is successful, regardless of the cause for the handover and this metric is calculated by

dividing successful handovers to the same or alternative gNB by handover attempts to the same or alternative gNB.

(2) Mean time of Inter-gNB handover execution of network slice: defines the successful mean time of Inter-gNB handover for a network slice and is used to measure network slice usage and performance and is determined by recording the time between the Source NG-RAN receiving a ”Release Resource” message from the Target NG-RAN and sending an ”N2 Path Switch Request” message to the Target NG-RAN over a specified period.

(3) Successful rate of mobility registration updates of Single network slice: This KPI reflects the rate of successful mobility registered changes in a specific network slice at the AMF and is calculated by dividing the number of successful mobility registration update requests received by the AMFs of a particular network slice by the number of successful mobility registration update requests received.

(4) 5GS to EPS handover success rate: A KPI that indicates the frequency with which a handover from 5GS to EPS is successful, regardless of the cause for the handover and is calculated by dividing the total amount of successful handovers from 5GS to EPS by the total amount of handover attempts from 5GS to EPS.

5.3.6 Energy Efficiency (EE) KPI

This metric indicates mobile service data energy efficiency in NG-RAN that is operating and is computed by dividing the Data Volume (DV) of the investigated network nodes by their Energy Consumption (EC). The detail of the formula could be found in [3] section 6.7.

6 Conclusion

In this report, we briefly introduced the fundamentals of the 5G network, common architecture and some use cases to describe how its components work. In addition, we covered Network Slicing, the key technology to enable the 5G potential. Finally, we explain service assurance in 5G networks based on QoS concept and E2E network KPIs, which assist to monitor and improve performance in 5G networks.

References

[1] *System Architecture for the 5G System V16.14.0*. 3GPP, September 2022.

[2] *ITU-T Recommendation E.800: "Definitions of terms related to quality of service"*. International Telecommunication Union, September 2008.

[3] *5G; Management and orchestration; 5G end to end Key Performance Indicators (KPI) (3GPP TS 28.554 version 16.7.0 Release 16)*.

[4] *O-RAN Working Group 6 (Cloudification and Orchestration) Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN v4.00 TR*. O-RAN Alliance e.V, 2022.

[5] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: challenges and research advances," *IEEE Network*, vol. 28, no. 6, pp. 6–11, 2014.

[6] X. Li, M. Samaka, H. A. Chan, D. Bhamare, L. Gupta, C. Guo, and R. Jain, "Network slicing for 5G: Challenges and opportunities," *IEEE Internet Computing*, vol. 21, no. 5, pp. 20–27, 2017.

[7] N. Al-Falahy and O. Y. Alani, "Technologies for 5G networks: Challenges and opportunities," *IT Professional*, vol. 19, no. 1, pp. 12–20, 2017.

[8] Y. Zong, C. Feng, Y. Guan, Y. Liu, and L. Guo, "Virtual network embedding for multi-domain heterogeneous converged optical networks: Issues and challenges," *Sensors*, vol. 20, no. September, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/9/2655

[9] "What is a 5G transport network?" September 2021. [Online]. Available: https://info.support.huawei.com/infofinder/encyclopedia/en/5G+Transport+Network.html

[10] I. Da Silva, S. E. El Ayoubi, O. M. Boldi, Ö. Bulakci, P. Spapis, M. Schellmann, J. F. Monserrat, T. Rosowski, G. Zimmermann, D. Telekom *et al.*, "5G ran architecture and functional design," *METIS II white paper*, 2016.

[11] *Functional Splits: the foundation of an Open 5G RAN.* Olli Andersson, Benetel, May 2021. [Online]. Available: https://www.5gtechnologyworld.com/functional-splits-the-foundation-of-an-open-5g-ran/

[12] *Open RAN functional splits, explained.* Eugina Jordan, Parallel Wireless, February 2021. [Online]. Available: https://www.5gtechnologyworld.com/open-ran-functional-splits-explained/

[13] I. Pupo, A. Santoyo-Gonzalez, and C. Cervelló-Pastor, "A framework for the joint placement of edge service infrastructure and user plane functions for 5G," *Sensors*, vol. 19, September 2019.

[14] *5G Mobile Networks: A Systems Approach (Synthesis Lectures on Network Systems).* Morgan & Claypool, July 2020.

[15] *5G security - enabling a trustworthy 5G system.* Ericsson, March 2018, [Online; accessed 2022-11-08].

[16] *5G Network Slicing Self-Management White Paper.* Huawei, October 2020. [Online]. Available: https://www-file.huawei.com/-/media/corporate/pdf/news/5g-network-slicing-self-management-white-paper.pdf

[17] D. Bega, M. Gramaglia, A. Garcia-Saavedra, M. Fiore, A. Banchs, and X. Costa-Perez, "Network slicing meets artificial intelligence: An ai-based framework for slice management," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 32–38, 2020.

[18] M. Li, J. Gao, C. Zhou, W. Zhuang *et al.*, "Slicing-based ai service provisioning on network edge," *arXiv preprint arXiv:2105.07052*, 2021.

[19] C.-N. Nhu and M. Park, "Dynamic network slice scaling assisted by attention-based prediction in 5G core network," *IEEE Access*, vol. 10, pp. 72 955–72 972, 2022.

[20] E. Coronado, R. Behravesh, T. Subramanya, A. Fernández-Fernández, S. Siddiqui, X. Costa-Pérez, and R. Riggio, "Zero touch management: A survey of network automation solutions for 5G and 6G networks," *IEEE Communications Surveys & Tutorials*, 2022.

[21] *5G Service requirements for the 5G system (3GPP TS 22.261 version 16.14.0 Release 16)*. 3GPP, April 2021.

Abbreviations

3GPP	The 3rd Generation Partnership Project
5G	The fifth-generation technological standard
5GS	5G system
5QI	5G QoS Identifier
AES	Advanced Encryption Standard
AI	Artificial intelligence
AMF	Mobility Management Function
AN	Access network
API	Application programming interface
AR	Augmented reality
ARP	Allocation and retention priority
AUSF	Authentication Server Function
BSS	Business support system
CN	Core network
CP	Control plane
CQI	Channel quality information
CSP	Communication service provider
CU	Central unit
CUPS	Control and user plane separation
DL	Downlink stream
DN	Data network
DNN	Data network name
DRA	Diameter Routing Agent
DRB	Data radio bearer
DSCP	Differentiated Services Code Point
DU	Distributed unit
E2E	End-to-end
eCPRI	Enhanced Common Public Radio Interface

EE	Engergy Efficiency
eMBB	Enhanced Mobile Broadband
ESP	Encapsulating security payload
EVPN	Ethernet VPN
FDD	Freuency division duplex
GBR	Guaranteed bit rate
GFBR	Guaranteed flow bit rate
GPRS	General Packet Radio Service
GTP	GPRS Tunnelling Protocol
I-UPF	Intermediate UPF
IoT	Internet of things
KPI	Key performance indicator
LI	Lawful Intercept
MAC	Medium access control
MEC	Multi-access edge computing
MFBR	Maximum flow bit rate
MIMO	Multiple input multiple output
ML	Machine learning
mMTTC	Ultra Reliable Low Latency Communications
MPLS	Multiprotocol Label Switching,
NEF	Network Exposure Function
NF	Network function
NFR	NF Repository Function
NFS	Network function services
NFV	Network function virtualization
NG-RAN	Next generation radio access network
NSSF	Network Slicing Selector Function
NR	New radio
O-RAN	Open radio access network

OSS	Operation support system
P2P	Point-to-point
PCF	Policy Control Function
PDCP	Packet Data Convergence Protocol
PDR	Packet detection rule
PDU	Protocol Data Unit
PFD	Packet flow description
PGW	Packet Data Network Gateway
PLMN	Public land mobile network
PSA	PDU session points
QFI	QoS flow identification (ID)
QoE	Quality of user experiences
QoS	Quality of services
RAN	Radio access network
RAT	Radio Access Technologies
RLC	Radio link control
RQA	Refective QoS Attribute
RRC	Radio Resource Control
RSA	Rivest-Shamir-Adleman encryption
RU	Radio unit
SA	Service assurance
SBA	Service-based architecture
SCTP	Stream Control Transmission Protocol
SDF	Service data flow
SDN	Software-defined networking
SDSF	Structured Data Storage Network Function
SGW-U	Serving Gateway User plane function
SLA	service level agreement
SMF	Session Management Function

SNSSAI	Session establishment success rate of one network slice
SQL	Structured Query Language
SRv6	Segment Routing IPv6
SST	Standardized slice type
TCO	Total cost of ownership
TN	Transport network
UDM	Unified Data Management
UDSF	Unstructured Data Storage Network Function
UE	User equipment
UL	Uplink stream
UP	User plane
UPF	User plane function
uRLLC	Massive Machine Type Communications
vEPC	Virtual evolved packet cores
VM	Virtual machine
VNF	Virtual network functions
VPN	Virtual Private Network
VR	Virtual reality
vRAN	Virtual radio access network